一种针对 BitTorrent 协议中 Have 消息的隐写分析方法 *

徐心怡，翟江涛†，戴跃伟

(江苏科技大学 电子信息学院，江苏 镇江 212003)

摘　要：网络隐写是一种以计算机网络通信数据为载体的隐蔽通信技术。BitTorrent 协议的巨大流量使其成为一种极佳的隐写载体，基于 BitTorrent 协议 Have 消息的隐写即在此背景下提出，目前公开文献尚无有效的检测算法。基于此，提出了一种基于多特征分类的检测方法。该方法首先提取正常 Have 消息数据流，接着提取均值、方差与直方图特征，最后基于 Adaboost 分类器给出了检测结果。实验结果表明，所提方法在观测窗口达到 1 000 个数据包时对该隐写的识别正确率可达 96%，在检测基于 Have 消息的隐写时具有良好效果。

关键词：BitTorrent 网络；网络隐写分析；信息安全

中图分类号：TP393　　doi: 10.3969/j.issn.1001-3695.2017.10.0939

Steganalysis method for Have message in BitTorrent protocol

Xu Xinyi, Zhai Jiangtao†, Dai Yuewei

(*School of Electronics & Information Jiangsu University of Science & Technology*, *Zhenjiang Jiangsu* 212003, *China*)

Abstract: Network steganography is a kind of computer network communication data as the carrier of the hidden communication technology. BitTorrent protocol of the huge flow to make it an excellent steganographic carrier, based on BitTorrent protocol has the message of steganography that in this context, the current public literature there is no effective detection algorithm. Based on this, this paper presented a multi-feature classification based on the detection method. Firstly, it extracted the method from the normal data stream, then it generated the secret data according to the concealed method, and extracted then the normal data and the dense data. Finally, it given the detection result based on the Adaboost classifier. The experimental results show that the accuracy of the proposed method can reach 96% when the observation window reaches 1000 packets, it has good effect in detecting steganography based on have messages.

Key Words: BitTorrent network, network steganalysis; information security

0　引言

随着信息技术的快速发展以及多种异构网络的融合，各行各业的不同业务逐步接入到互联网中，对网络的依赖越来越强，随之而来的安全问题越来越严重。传统加密技术是将可读的数据变成杂乱无章的密文，虽使信息不可读，但却暴露了通信的存在性。网络隐写[1,2]作为一种隐蔽通信方式，利用合法的数据流作为载体在网络中传递秘密信息，隐藏了秘密信息传输通道的存在，具有较高的隐蔽性。以高级持续性威胁(advance persistent threat, APT)为代表的网络攻击，为了提高自身的生存性，往往使用了隐蔽通信技术[3,4]。网络隐写作为隐蔽通信的一种，针对其的检测可为 APT 防御提供技术支持[5]，这也使得网络隐写检测成为近年来研究的热点。

在网络隐写研究中，由于 P2P[6]应用流量巨大，是一个较为理想的隐写载体，近年来提出了众多隐写方法，而针对其上隐写分析[7]则发展较为缓慢，为了实现针对 APT 的防御，迫切需要解决 P2P 中的隐写检测问题。本文针对 P2P 中典型的 BitTorrent 文件共享协议中 Have 消息的秘密信息传输方法进行分析，提出一种基于 Adaboost 模式识别的隐写检测方法。经实验验证表明，本文所提方法具有较高的检测准确率。

1　背景知识

隐写术将秘密信息嵌入到合法载体[8]中，隐藏信息或通信的存在性。常见的载体信道有图像、音频、视频和网络数据流等[9]。而网络隐信道是隐信道技术中的一种，它是利用公开信道上传输的合法数据包作为载体，在其中嵌入秘密消息的一种

基金项目：国家自然科学基金资助项目（61472188，61602247，61702235，U1636117）；江苏省自然科学基金资助项目（BK20150472，BK20160840）；CCF-启明星辰“鸿雁”科研基金资助项目（2016011）

作者简介：徐心怡（1993-），女，江苏盐城人，硕士研究生，主要研究方向为多媒体与信息安全；翟江涛（1983-），男（通信作者），河南三门峡人，副教授，主要研究方向为多媒体与信息安全（jiangtaozhai@gmail.com）；戴跃伟（1962-），男，江苏镇江人，教授，博导，主要研究方向为多媒体与信息安全、系统工程理论及应用、复杂系统管理控制.

隐蔽通信技术。1973 年，Lampson[10]首次提出了隐信道的概念，将其定义为：隐蔽信道是指利用资源共享创建的、违反系统本身意图且对系统安全造成危害的一种通信机制。Girling 在 1987 年首次提出网络隐蔽信道[11]，是指在现代网络通信中存在的危害网络安全的通信信道。在海量的网络数据流中，网络通信体现出随机性和动态性特点，并且网络隐信道可以绕开防火墙、入侵检测等安全设备，具有强隐蔽性特点，使攻击方难以跟踪取证。

P2P 网络是目前使用最为广泛的文件共享方式，BitTorrent (BT)是 P2P 的一种典型应用。基于 BT 协议的 P2P 网络数据流大体可分为 BT 种子文件、BT 服务器文件和 BT 消息文件三部分。基于 BT 种子文件的网络隐写方法主要是利用大小写不敏感变换和结构冗余复用技术，将秘密信息嵌入到 BT 种子文件各种关键字的冗余空间中[12]。BT 服务器，又称之为 Tacker 服务器，用于保存 BT 种子文件以及记录当前下载者的网络信息。基于 Tacker 服务器的网络隐写方法主要分为两种：一种是利用 HTTP GET 请求消息中关键字 peer_id 的冗余空间为载体嵌入秘密信息[13]，另一种是通过 HTTP 消息将秘密信息直接写入 Tacker 服务器[14]。基于 BT 消息的网络隐写方法主要有文献[15]提出的基于 Bitfield 消息的信息隐藏算法和基于 Piece 消息的信息隐藏算法则是将秘密信息嵌入到 Bitfield 消息以及 Piece 消息的冗余空间中。其中，基于 Bitfield 消息的信息隐藏算法会引入当前 P2P 节点所拥有数据块数目异常，导致频繁出现来至其他 P2P 节点的请求消息；基于 Piece 消息的信息隐藏算法会引入当前 P2P 节点所拥有数据块内容异常，导致频繁出现来至其他 P2P 节点的重传消息。因此，这两种方法都会主动地给正常的 P2P 通信带来出错异常，隐蔽性较差，而且这两种方法易受到网络复杂环境的干扰，鲁棒性较差。

Peer 消息作为 BitTorrent 网络通信不可或缺的组成部分，对于整个 BT 网络的运行起着至关重要的作用，而且 Peer 消息通信容量大。Have 消息是一种通信频繁且分布有序的 Peer 消息[16]，每当一个完整的数据块下载完毕，该客户端节点 Peer 即可向外发送与之对应的 Have 消息。由于 Have 消息具有通信频繁、分布有序以及修改负载不会引起系统错误的特点，使得其具备了承载网络隐蔽通信的冗余空间。

由于 Have 消息是 BT 消息文件中有一个具备向已连接节点宣称自己拥有某个数据块功能的消息，数据块大小固定且共享文件越大，共享文件数据块的个数则越多，相应的 Have 消息的个数也就越多，所以若以 Have 消息序列为载体嵌入秘密信息，嵌入容量足够大。除此之外，由于数据块 Piece 下载的随机性，与之对应的 Have 消息索引号也是随机的，若将秘密信息嵌入到 Have 消息的排序中，使秘密信息随 Have 消息一起进行传输，这种方法的隐蔽性较强。由于引入了信息校验机制，即便 Have 消息受到网络复杂环境的干扰而出错，也会通过该校验机制保证数据传输的准确性，鲁棒性也相对较好。

近年来 P2P 技术飞速发展，基于 P2P 的隐蔽通信方法越来越多，其中基于 Have 消息的隐写方法具有较高的隐蔽性与鲁棒性。而现有的网络信息隐藏检测技术的研究主要集中于使用 TCP/IP 协议的 Internet 网络，对于使用 BT 协议的 P2P 网络的研究相对较少，且尚无公开文献提出对于 Have 消息隐藏通信的检测方法。因此，以 P2P 网络数据流作为载体的信息隐藏检测技术值得进一步地深入研究，本文针对 Have 消息的隐蔽通信提出了一种检测方法。

2　基于 BT 协议 Have 消息排序的信息隐藏方法

文献[17]提出了一种基于 BitTorrent 协议 Have 消息的秘密信息传输方法。其使用的隐藏算法原理是通过改进奇偶映射信息编码方式来实现信息隐藏算法的功能。原有奇偶映射方法的前提是不改变子数组中各元素的值，仅依靠元素自身的奇偶性和排序来实施编码。而这种方法中改进后的奇偶映射编码则是打破这个前提，它会利用“加 1 或减 1”的方式改变部分元素的奇偶性，以此来提高载体利用率，即利用 N 个原始数据负载了 N b 的秘密信息，改进后的载体利用率 R_2 为

$$R_2 = N / N * 100\% = 100\% \tag{1}$$

这种方式虽然改变了子数组中部分元素值，但改变后的元素值依然属于整个数组集合，并不会发生异常，其效果相当于是在整个数组集合中改变各元素的顺序。基于改进后的奇偶映射信息编码方式，若待嵌入的比特流信息集合为:B = {b1，b2，…，bi，…，bN}，其中，bi= ‘0’ 或 bi= ‘1’ 且 1 ⩽ i ⩽ N，并结合式(2) 的 Have 消息序列，编码所得新序列值为

$$h_i = \begin{cases} h_i + 1, & if\ h_i = 0 且 b_i = '1' \\ h_i, & if \begin{cases} h_i\%2 = 0 且 b_i = '0' \\ h_i\%2 \neq 0 且 b_i = '1' \end{cases} \\ h_i - 1, if\ h_i \neq 0 且 & \begin{cases} h_i\%2 = 0 且 b_i = '1' \\ h_i\%2 \neq 0 且 b_i = '0' \end{cases} \end{cases} \tag{2}$$

尚无公开文献提出对于该 Have 消息隐蔽通信的检测方法。

3　本文算法

3.1　特征提取

本文提取正常通信与隐蔽通信的均值、方差、直方图三种特征。均值是表示一组数据集中趋势的量数，它是反映数据集中趋势的一项指标，它既可以用来反映一组数据的一般情况和平均水平，也可以用它进行不同组数据的比较，以看出组与组之间的差别。用均值表示一组数据的情况，有直观、简明的特点。方差是在概率论和统计方差衡量随机变量或一组数据时离散程度的度量。概率论中方差用来度量随机变量与其数学期望之间的偏离程度。统计中的方差是每个样本值与全体样本值的平均数之差的平方值的平均数。在统计工作中，均值和方差是描述数据资料集中趋势和离散程度的两个最重要的测度值。为了进一步提高分类识别的准确率，本文引入直方图特征来更好

地描述数据流的分布，并基于该统计方法建立了正常数据流与含密数据流在 0～N 上的的分布概率统计模型。

3.2 Adaboost 分类器

Adaboost 作为 Boosting 算法中的典型一种，通过迭代，能够对同一训练集训练多个弱分类器，并通过一定的方式将这些弱分类器组合起来，提升为一个更强的最终分类器。在训练过程中，算法根据每次弱分类器的分类正确与否，以及上次总体分类准确率来确定每一个样本的权重。对于误检测的样本，其权重也就越大，这就相当于改变了数据的分布形式。对修改的新样本数据进行下一层的弱分类器训练，迭代多次达到收敛条件以后将每次训练得到的分类器组合起来，就得到了最终的决策分类器。标准的 Adaboost 分类器是一种线性加权形式的集成分类器，即

$$f(x)=sgn\left(\sum_{t=1}^{T}\alpha^{(t)}h^{(t)}(x)\right) \tag{3}$$

其中：强分类器 $f(x)$ 是 T 个弱分类器 $h^{(t)}(x)$ 的线性组合，弱分类器加权系数 $\alpha^{(t)}$ 是由 Adaboost 算法训练得到。Adaboost 算法具有非常显著的优点：

a)Adaboost 是一种具有高分类精度的分类器；

b)Adaboost 提供的是一种提升框架，其中的弱分类器可以有多种形式；

c)不用担心过拟合问题，泛化性能好。

基于 Adaboost 算法的这些显著特点，本文采用 Adaboost 算法进行分类器的训练，其中的弱分类器本文使用决策树。

4 实验结果及分析

4.1 数据预处理

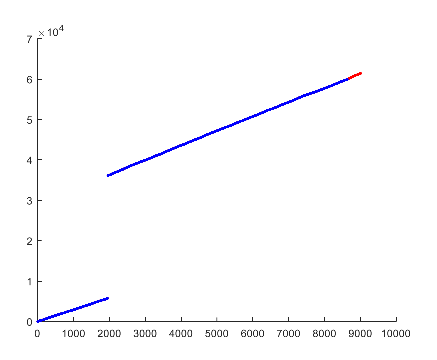
如图 1(a)所示，蓝色数据为含秘数据，红色数据为正常数据（见电子版），由于数据太过接近，所以在图中有所重叠；图(b)为两段数据之差，能够看出两者差异仅为 0 或 1，远远小于数据流数值大小，此时难以体现数据流之间的差异性。因此要对正常数据与含密数据进行分离，有必要增大数据流之间的差异性，一种有效的办法就是对数据流的前后数据求差。图 2(a)即为正常数据前后差值图。结合图 2(b)能够看出通过差值，数据流基本限制在-20～20 间，因此极大地增加了 0 或 1 差异对正常数据与含密数据的差异影响。但由于数据流在某些时候具有极大的跳变，导致数据流前后差值在某些时候数值过大，影响了进一步的数据分离。

为了进一步增加数据流之间的差异，减小跳变数据的影响，本文将差值数据归一化到 0～N-1 间，其中 N 为一个较小的正整数，其方法就是用差值数据流对 N 取余。

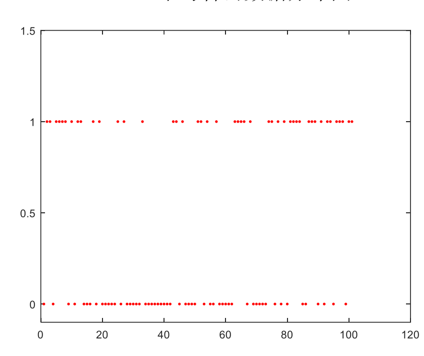
结合如上所述的数据流前后差值与取余两步操作，便极大增加了正常数据与含密数据之间的差异性，其数学描述如式（4）所示。

$$X^{'}=\mathrm{mod}(\Delta(X),N) \tag{4}$$

其中：$\Delta(\cdot)$ 为差值函数，$\mathrm{mod}(\cdot,N)$ 是对 N 的取余函数。

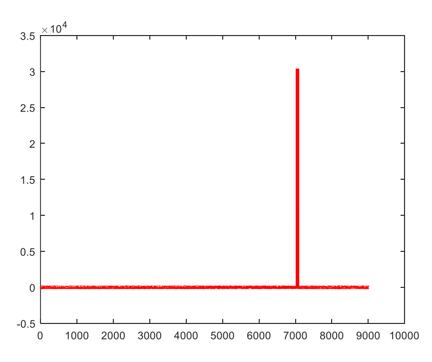


（a）正常与含密数据分布图

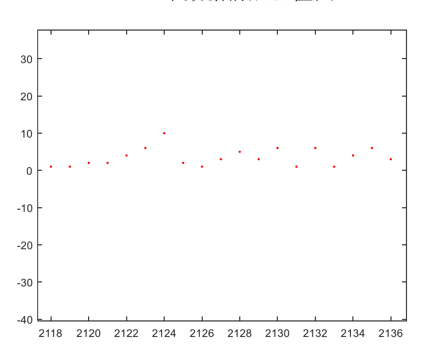


（b）正常与含密数据差值图

图 1 正常与含密数据示意图



（a）正常数据前后差值图



（b）正常数据前后差值局部示意图

图 2 正常数据前后数据差值图

4.2 特征提取

数据流的均值与方差能够体现数据的整体分布情况。图 3 所示即为正常数据与含密数据均值和方差的分布图。能够看出，数据流的均值与方差分布仅有小部分的重叠，说明数据流分布对正常数据与含密数据分类识别的有效性。为了进一步提高分类识别的准确率，本文引入直方图特征来更好地描述数据流的分布，其基本思想就是统计数据流在 0～N 上的分布概率。因此对于经过预处理的数据流 X'，特征提取过程如式(5)所示。

$$F(X')=[mean(X'),\mathrm{var}(X'),hist(X')] \tag{5}$$

其中：$mean(\cdot)$ 为均值函数；$\mathrm{var}(\cdot)$ 为方差函数；$hist(\cdot)$ 为直方图函数。

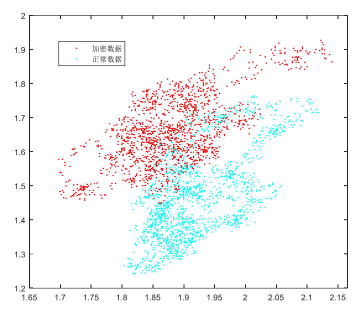


图 3　正常数据与含密数据均值方差分布图

4.3 检测结果

实验截取原始数据流，按照每个窗口包含 w 个信息对其进行划分，嵌入隐秘信息之后，按照同样方法进行划分，这样就分别得到了用于分类识别的正常及含密数据，其中 70%用于训练，30%用于测试。按照第 3 章中介绍的方法，计算出不同窗口下的均值、方差及直方图特征，使用 Adaboost 算法进行分类器的训练与测试。下面实验分析窗口大小 w、直方图特征量化等级 N 以及最终特征维数 k 对识别率的影响。

1)窗口大小 w 对识别率的影响

为了分析窗口大小 w 对识别率的影响，首先固定 N=5，k=2，研究 w 在取值 200、400、600、800、1 000、1 200、1 400、1 600、1 800 以及 2 000 这 10 种情况下的识别率。图 4 即为窗口大小 w 对识别率的影响结果。从中能够看出，窗口大小较大影响了最终的识别效果，这主要是因为更多的数据量能够更为准确地挖掘数据的内部特性，找到正常数据与含密数据之间的差异性。

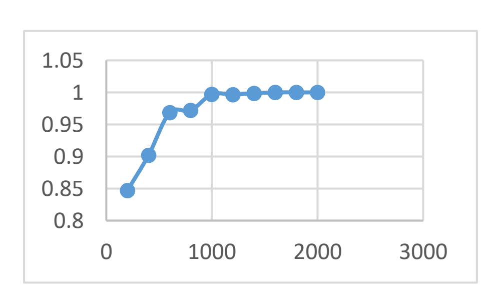


图 4　窗口大小 w 对识别率的影响结果图

2)直方图特征量化等级 N 对识别率的影响

为了分析直方图特征量化等级 N 对识别率的影响，首先固定 w=1000，k=2，研究 N 的取值对识别率的影响。直方图特征量化等级 N 对识别率的影响结果如图 5 所示。从图 5 能够看出，N=5 时识别效果最好，当量化等级较小时，直方图特征难以准确描述数据的分布特点；而当量化等级较大时，虽然能够描述数据分布特点，但直方图特征之间包含了更多的冗余信息，会较大影响分类器的泛化能力，因此量化等级过小或者过大都不利于最终的识别率。

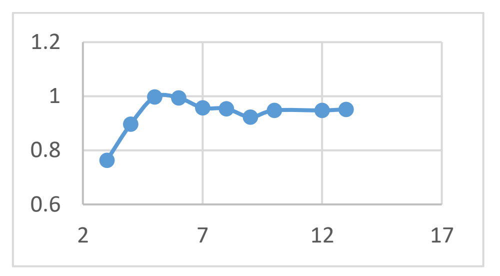


图 5　直方图特征量化等级 N 对识别率的影响结果

5 结束语

本文对 BitTorrent 协议中 Have 消息的隐写方法的算法原理和具体算法进行了研究分析，找到了这种新型隐写方式的特点。根据其特点，将模式识别的概念用于 Have 消息隐写的分析中，通过均值方差直方图三个特征来区分正常数据和含密数据。实验表明，当观测窗口大于 1 000 时，特征量化等级 N=5 ，特征维数 k=3 时，运用本文提出的方法区分正常数据和含密数据，可以达到最优的效果。

参考文献：

[1] Fridrich J. 数字媒体中的隐写术 [M]. 张涛，奚玲，张彦等，译. 北京：国防工业出版社, 2014: 2-9.

[2] Wendzel S, Zander S, Fechner B, et al. Pattern-based survey and categorization of network covert channel techniques [J]. ACM Computing Surveys, 2015, 47 (3): 50.

[3] Elzbieta Z, Wojcjech M, Krzysztof S, et al. Trends insteganography [J]. Communications of the ACM, 2014, 57 (3): 86-95.

[4] Rios R, Onieva J A, Lopez J. Covert communications throughnetwork configuration messages [J]. Computers & Security, 2013, 39 (4): 34-46.

[5] Kiyavash N, Koushanfar F, Colemant P, et al. A timingchannel spyware for the CSMA//CA protocol [J]. IEEE Trans onInformation Forensics & Security, 2013, 8 (3): 477-487.

[6] 吕晓鹏，王文东，龚向阳，等. 混合网中的 P2P 资源共享机制 [J]. 北京邮电大学学报, 2011, 34 (4): 113-117.

[7] 徐钒文. 基于 P2P 的隐蔽匿名通信技术研究 [D]. 北京：北京邮电大学, 2012: 4-51.

[8] Cunche M, Kaafar M A, Boreli R. Asynchronous covert communication using BitTorrent trackers [C]// Proc of International Conference on High

Performance Computing and Communications. Piscataway, NJ: IEEE Press, 2014: 213-291.

[9] Praviya B B S, Priyanka S S, Thamarai S V. Hiding of data using steganography technique [J]. International Journal of Engineering Sciences & Research Technology, 2015, 4 (2): 78-81.

[10] Lampson B W. A note on the confinemengt problem [J]. Communications of the ACM, 1973, 16 (10): 613-615.

[11] Desimine J, Johnson D, Yuan B, et al. Covert channel in the Bit Torrent tracker protocol [C]// Proc of International Conference on Security and Management. New York: Rochester Institute of Technology, 2012: 223-226.

[12] 李自帅, 孙兴明, 王宝威, 等. 一种对等网中的隐写方案 [C]// 智能信息隐藏与多媒体信号处理国际会议. 2008: 20-24.

[13] Desimone J, Johnson D, Yuan B. Covert channel in the BitTorrent tracker protocol [EB/OL]. (2012) http://scholarworks. rit. edu/other/300.

[14] Cunche M, Kaafar M, Boreli R. Asynchronous covert communication using BitTorrent trackers [C]// Proc of the 11th IEEE International Conference on Embedded Software and System, 2014: 827-830.

[15] 李子帅. 基于 BitTorrent 网络的信息隐藏技术研究 [D]. 长沙: 湖南大学, 2009.

[16] Zhang Lihua, Liu Guangjie, Zhai Jiangtao, et al. Improving reliability of covert timing channel to packet loss [J]. Journal of Information Hiding & Multimedia Signal Processing, 2015, 6 (3): 544-553.

[17] 高斌, 翟江涛, 戴跃伟. 基于 Bit Torrent 协议 Have 消息的信息隐藏方法 [J]. 计算机应用, 2017, 37 (1): 200-205.